Office Note No. 1

Joint Numerical Weather Prediction Unit

Preliminary Statistical Examination of Comparable Results From Two Numerical Forecasting Models

Louis P. Carstensen
U. S. Weather Bureau

31 March 1955

As the JNWP Unit approaches an operational stage it is desirable to take stock of the performance level of the several numerical forecast models presently available for its operations. The number and variety of situations for which forecasts have been made are too few to permit many definite conclusions; however, certain statistical summaries of the material that is available appear to be desirable and, for some purposes, adequate. Direct comparisons of different methods in identical situations are undertaken wherever possible.

Comparison with Subjective Forecasts

One comparison which was made was between 24-hour thermotropic forecasts of the 500 mb level and subjectively prepared 24-hour forecasts for the same pressure surface. The thermotropic charts were taken from a series of 60 24-hour forecasts made at G.R.D. at 12 hourly intervals through January 1953. The subjectively prepared charts which were used had been prepared on a routine basis on acetate work sheets but were not ordinarily made into a permanent record. It is thought that they were carefully prepared by skilled meteorologists. The 16 cases used in this comparison were all that were available. They were separated by time intervals of at least 24 hours.

This comparison was made on the basis of 24-hour changes of pressure gradient measured between two grid units (total distance about 5° of latitude). The computed and the subjectively forecast changes were each correlated with the observed changes of gradient. The east-west and the north-south components were grouped together so that 32 pairs of numbers were available for each point. Gradients were compared at

30 points distributed uniformly approximately between 75°W and 110°W and 39°N and 53°N. The correlations between computed and observed changes of gradient were plotted and the geographical distribution of the corre-coefficient is shown on Fig. 1.

The subjectively forecast gradient changes showed the higher correlation at 3 points on the edges of the area considered; elsewhere, the numerically computed gradient changes showed the higher correlations, the geographical distribution of the difference is shown in Fig. 2. Both forecast methods showed the higher skill in the east of the area and lesser skill in the west. The average of the correlation coefficients were .70 for the numerical and .56 for the subjective.

A valid criticism of this comparison is that probably the initial charts used for the numerical forecasts were more carefully analyzed than were the initial charts used by the subjective forecasters. There does not appear to be any information which could be used to assess quantitatively the importance of this difference of quality of initial maps. This objection is partially met by restricting the comparison to a region where the observations specify the analysis to within relatively small limits, but it may still be important in certain rare cases.

Correlation coefficients may be thought of as measuring the phase difference of the quantities being considered. Another item of interest is the relative magnitude of the quantities being compared. The measure of relative magnitude used here has been the standard deviations of the 24-hour changes. ~~The~~ ratio of the standard deviation of the numerically forecast gradient changes to that of the observed gradient changes was determined for each point. The average of these

30 ratios was 1.17 showing that for these 16 cases the amplitude of the computed forecast gradients tended to be slightly above that of the observed gradients. This tendency toward slight intensification seems to have been present, on the average, for results from all of the numerical forecast models examined up to this time.

Comparison between 3-level and Thermotropic Models

Through the cooperation of G.R.D. and I.A.S. there were seven cases available in which 24-hour forecasts have been made with both models for the same 24-hour interval and for approximately the same area. These cases were all wintertime situations and almost all involved development of a low pressure system. For the area common to both forecasts, correlation coefficients between forecast and observed 24-hour height changes were computed for each level of each forecast. Usually between 80 and 100 grid points could be used in these correlations. The results are given in table 1.

Table 1. Correlations between predicted and observed height changes for the same time interval and for a common area.

time	model level	900	700	400	850	P 500	200	thermotropic 1000	thermotropic 500
1500Z	24/11/50	.37	.73	.61	.62	.91	.25	.00	.14
0300Z	13/1/53	.54	.00	.23				.14	.75
0300Z	15/1/53	.71	.62	.42				.80	.77
0300Z	24/1/53	.76	.79	.88				.86	.96
0300Z	27/1/53	.85	.80	.84				.86	.89
1500Z	5/11/53	.49	.83	.88	.61	.75	.74	.55	.57
1500Z	13/3/54	.75	.89	.84	.78	.82	.61	.83	.89
mean r		.64	.67	.67				.58	.71

These results do not indicate any significant difference in skill between the models. It may be remarked in this connection that subjective evaluations of the forecast height changes and of the prognostic charts also failed to disclose any consistent superiority of one model over the other for this set of cases.

The significance of these results for the purpose of direct comparison is somewhat uncertain for several reasons in addition to the fact that the number of cases was small. Among the important differences in addition to the forecast model are the unequal size of the forecast areas, quite different boundary conditions, and some differences in initial conditions due to the circumstance that the models use different pressure levels. The differences in boundary conditions assumed or the adverse effects of either type of boundary conditions on forecast areas of the size treated may alone be sufficient to obscure any real difference in the performance of the models.

It does appear to be safe to conclude that the difference between the models is small compared to the error made with either model.

It is important not to take these correlation coefficients out of context. Because of the highly selected nature of the cases considered the coefficients are of value only for purposes of this comparison. An example of the selective nature of the cases is that several of the cases were chosen on the basis that one or the other model had performed relatively poorly on them.

<u>General Level of Forecast Quality</u>

Various statistical summaries of the series of 60 thermotropic forecasts are being prepared at G.R.D. These may be expected to give a fairly reliable indication of the general level of performance of the thermotropic model during winter months. The correlation coefficients between 24-hour forecast and observed changes have been given as about .75 for the 500 mb surface and approximately .65 for the 1000 mb surface. (1) This was for a rectangular area on a Lambert Conformal chart bounded approximately by the corner points 29.5N 72W, 55.5N 58W, 55.5N 131W and 29.5N 118W. Three rows and columns exterior to this area had been dropped. Some aspects of the errors appear to have been systematic, suggesting that, if necessary, certain empirical corrections could be made.

No comparable comprehensive study of the performance of the 3-level model has as yet been undertaken. However, an attempt has been made to summarize statistically all of the 3-level forecasts available. Fourteen forecasts using the 900-700-400 levels and eight forecasts using the 850-500-200 levels have been collected mainly through the cooperation of I.A.S. These forecasts were not all for the same geographical area but they all overlap to some extent. Forecast height changes and observed height changes were determined for as many as possible of a set of 100 grid points in fixed geographical locations. These points were spaced 300 km apart each way in rows parallel to the 85th meridian with the southwest corner of the grid at approximately 97W 30.3N.

Correlation coefficients were computed for each level for each forecast and are listed in table II.

Table II. Correlations between predicted and observed height changes for that part of the forecast which fell on a fixed "verification" grid.

				σ			p	
Hour	date	model level	900	700	400	850	500	200
0300Z	23/11/50					.82	.89	.57
1500Z	23/11/50					.66	.72	.88
0300Z	24/11/50		.47	.87	.82	.71	.88	.67
1500Z	24/11/50		.37	.73	.61	.62	.91	.25
0300Z	25/11/50					.62	.75	.50
0300Z	24/11/52		.91	.88	.20			
1500Z	24/11/52		.83	.86	.91			
0300Z	25/11/52		.78	.72	.83			
1500Z	26/11/52		.97	.95	.96			
0300Z	13/1/53		.78	.03	.17			
0300Z	15/1/53		.42	.34	.61			
0300Z	24/1/53		.76	.79	.88			
0300Z	27/1/53		.84	.79	.82			
1500Z	14/4/53		.70	.81	.72			
0300Z	15/4/53		.22	.87	.76			
1500Z	5/11/53		.42	.83	.88			
0300Z	6/11/53					.58	.86	.77
1500Z	13/3/54					.22	.67	.56
			.75	.89	.84	.78	.82	.61
mean r			.66	.74	.72	.63	.81	.60

These do not appear to differ significantly from the level of results obtained at G.R.D. with the aid of the thermotropic model. It should again be stressed that these figures are for a highly selected group of cases, a number of which involve "explosive" development. An additional reason for not regarding these coefficients as an absolute measure of skill is that the effects of removing "known" systematic errors have not been investigated. Perhaps the chief value of these coefficients is in comparing the accuracy of the prognostic charts at the various levels within one forecast model.

Variation of Performance Level with Geography

It was noted above that forecasts of gradient changes at the 500 mb level using both the thermotropic model and subjective methods were better in the southeast and less accurate in the west and north of the forecast area. Likewise the correlations of observed 24-hour height changes of the 500 mb level with those predicted by the thermotropic model were highest in the southeast with a secondary maximum in the southwest and were lowest in the northwest and through the central plains. A well-defined center of low correlation values in the central plains (Kansas and Oklahoma) was also evident for the 1000 mb forecast made using the thermotropic model. It appears to be quite certain that this center of low skill results from a physical effect of the Rocky Mountains on the flow, or to some extent results from the determinations of the height of the 1000 mb surface where pressure at the earth's surface is less than 1000 mb.

Correlation coefficients and constants of linear regression have been computed for the 14 available cases for 25 grid points distributed over the verification area for each level of the 900-700-400 forecasts. These coefficients which suggest a level and distribution of performance very similar to that exhibited by the thermotropic model are plotted in their appropriate positions in Fig. 3.

The verification area does not extend far enough to the west to indicate whether a secondary maximum performance area such as that exhibited by the thermotropic model would be found over the southwest. The correlation coefficients do appear to suggest that an area of low skill in the lower levels is to be found somewhat to the east of the great plains' center of low skill exhibited by the thermotropic model.

<u>Systematic Height Errors</u>

Although systematic errors were not the subject of this study, it was convenient to look at one aspect of such errors as a by-product of the correlation statistics.

Subjective examination of forecasts prepared with the aid of the 3-level model has occasionally suggested that false anticyclogenesis may occur especially to the southwest of the low center. One method of investigating this is to examine the constant "a" of the linear regression equation

$$\Delta z_{obs} = a + b \Delta z_{fcst}$$

These constants are plotted in appropriate geographical positions in Fig. 3. The number of cases going into the determination of a constant varies from 14 near the center of the verification area to approximately 8 on the boundaries.

Although much irregularity due to the small and variable sample is evident, the constant "a" does appear to vary at each level from -50' or -100' in the south central portion of the grid to 50' to 200' on the north, east and western boundaries of the verification area. This result supports the general form of the expected systematic errors of pressure, but suggests an arrangement of average height error which differs somewhat from the prevalent subjective impression.

(1) G.R.D. Report August 1954

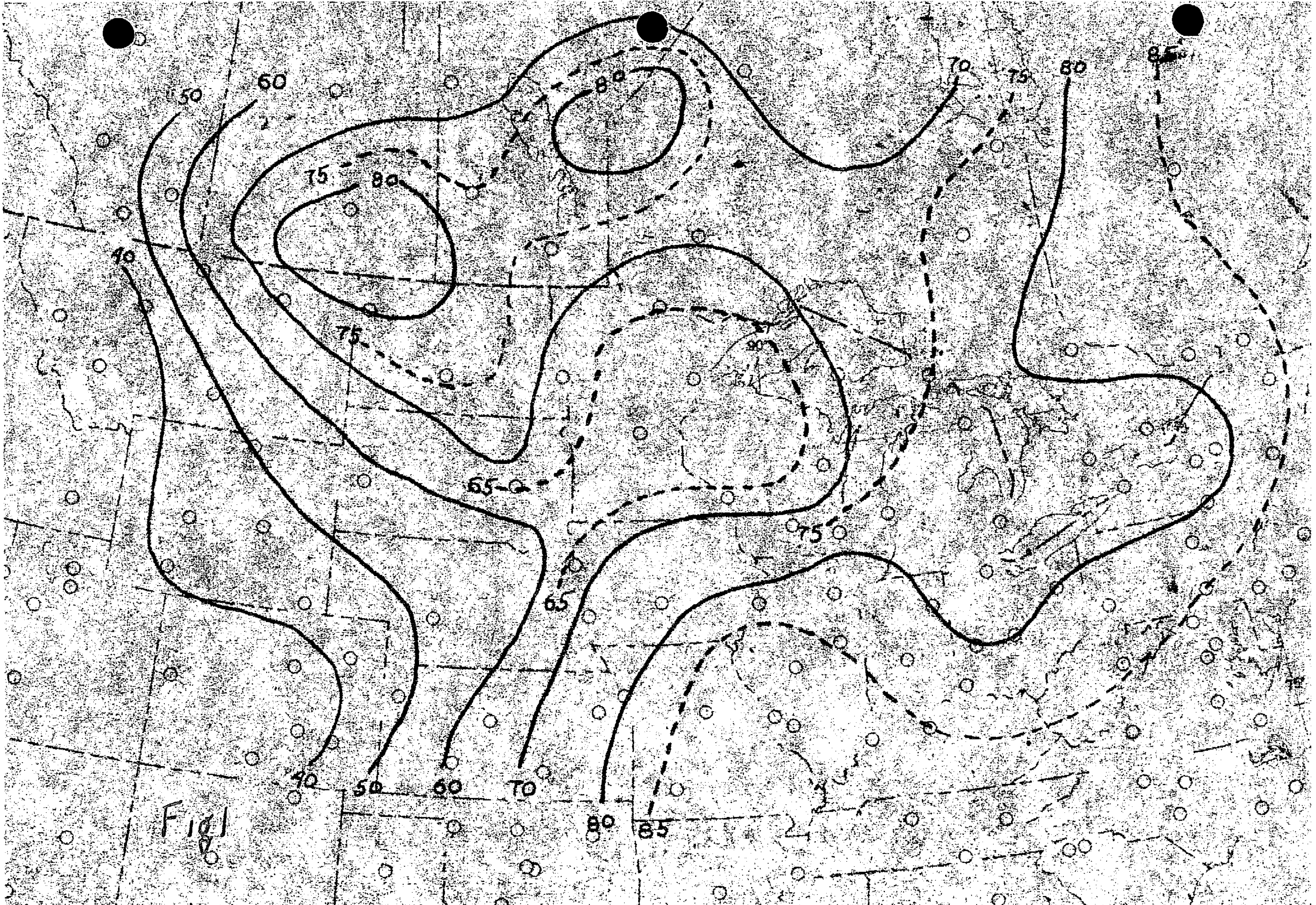

Fig 1

Correlation — Observed Vs Forecast
24 Hour Changes In Height Gradient
Numerical
$\bar{r}$ = .701

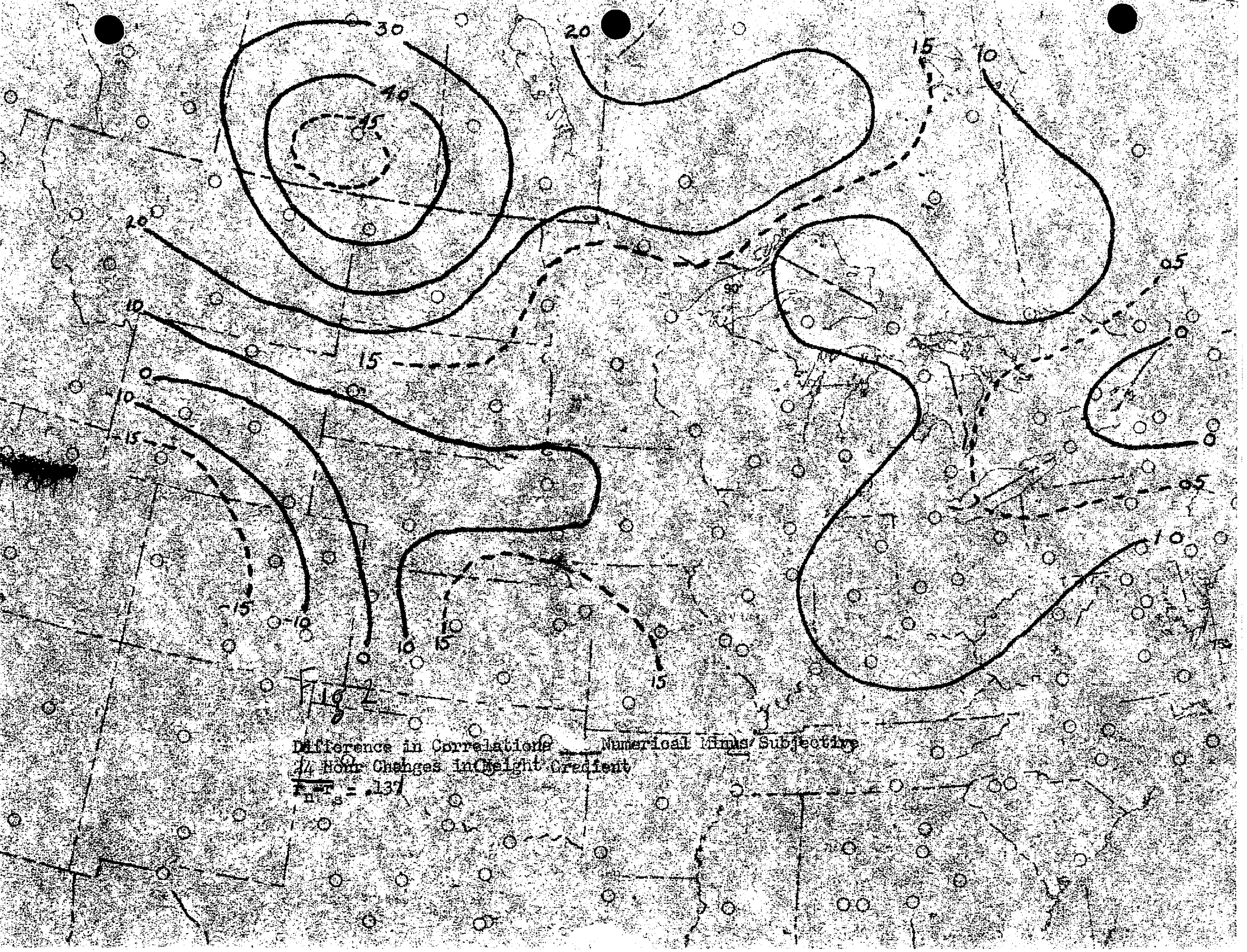

Fig 2

Difference in Correlations — Numerical Minus Subjective
24 Hour Changes in Height Gradient
$\overline{r_n - r_s} = .137$

Summary of 14 forecasts
9-7-4 Model
Arrangement of data
4 r a b
7 r a b
9 r a b
r = Correlation Coefficient
a&b for liniar regressions relationship
$\Delta Z_o = a(\text{in feet}) + bZ_f$
ΔZ_o = Verification ΔZ value
ΔZ_f = Forecast ΔZ value

.17 55 .08 .29 101 .20	.78 59 .47 .96 98 .25	.76 73 .56 .45 110 .29	.76 109 .53 .64 156 .94	.81 78 .78 .65 169 .57
.53 172 .46 .35 71 .18 .32 136 .19	.82 186 .65 .83 101 .69 .56 161 .51	.83 213 .79 .85 76 .68 .68 90 .49	.89 152 .78 .85 101 .64 .86 87 .45	.80 127 .80 .86 141 .74 .83 156 .55
.76 248 .90 .42 108 .46 .49 163 .52	.88 127 .95 .60 -28 .64 .32 44 .38	.85 84 .72 .79 -20 .59 .53 -28 .31	.80 166 .63 .89 76 .61 .70 -9 .38	.82 213 .58 .88 80 .71 .90 118 .60
.74 178 .85 .56 80 .59 .72 131 .75	.79 68 .77 .27 -98 .23 .48 50 .48	.87 -8 .73 .72 -99 .47 -.07 -162 -.03	.83 15 .59 .91 95 .89 .69 -19 .49	.95 128 .48 .94 -19 .76 .89 61 .63
.36 139 .37 .88 99 .99 .52 98 .48	.55 121 .38 .30 -4 .30 .47 34 .33	.63 -73 .53 .46 -77 .37 -.21 -37 -.12	.60 -68 .59 .65 46 .91 -.06 -181 -.05	.82 -14 .48 .79 -20 .77

Fig 3